Rademacher Complexity of Margin Multi-category Classifiers

Yann Guermeur

Noname manuscript No.
(will be inserted by the editor)

Rademacher Complexity of Margin Multi-category Classifiers

Yann Guermeur

Received: date / Accepted: date

Abstract One of the main open problems of the theory of margin multi-category pattern classification is the characterization of the optimal dependence of the confidence interval of a guaranteed risk on the three basic parameters which are the sample size m, the number C of categories and the scale parameter γ. This is especially the case when working under minimal learnability hypotheses. The starting point is a basic supremum inequality whose capacity measure depends on the choice of the margin loss function. Then, transitions are made, from capacity measure to capacity measure. At some level, a structural result performs the transition from the multi-class case to the bi-class one. In this article, we highlight the advantages and drawbacks inherent to the three major options for this decomposition: using Rademacher complexities, covering numbers or scale-sensitive combinatorial dimensions.

Keywords margin multi-category classifiers; Rademacher complexity; metric entropy; margin Graph dimension

Mathematics Subject Classification (2000) 68Q32 · 62H30

1 Introduction

In the framework of agnostic learning, one of the main open problems of the theory of margin multi-category pattern classification is the characterization of the way the confidence interval of a guaranteed risk should vary as a function of the three basic parameters which are the sample size m, the number C of categories and the scale parameter γ (see [17] for a survey). This is especially the case when working under minimal learnability hypotheses. In

Yann Guermeur
LORIA-CNRS
Campus Scientifique, BP 239
54506 Vandœuvre-lès-Nancy Cedex, France
E-mail: Yann.Guermeur@loria.fr

that context, several basic supremum inequalities are available, for the different margin loss functions commonly used. They constitute starting points to derive bounds whose capacity measure is a Rademacher complexity, a metric entropy or a scale-sensitive combinatorial dimension. In the corresponding sequence of transitions, a specific step is a structural result performing the transition from the multi-class case to the bi-class one. It can involve any of the three kinds of capacity measures listed above. In this article, we study the incidence of the choice of the level for this decomposition when the margin loss function is either the parameterized truncated hinge loss or the margin indicator loss function. In the process of deriving the corresponding guaranteed risks, we establish a new combinatorial result involving covering numbers based on the L_∞-norm and a γ-Ψ-dimension: the margin Graph dimension. This capacity measure is bounded from above for the class of functions computed by the multi-class support vector machines (M-SVMs), which provides an illustrative example of the dependence on the three basic parameters of a confidence interval.

The organization of the paper is as follows. Section 2 deals with the theoretical framework and the margin multi-category classifiers, focusing on their capacity measures. Section 3 highlights the connections between these measures. It introduces the new combinatorial result. Sections 4 and 5 are devoted to two comparative studies: for the parameterized truncated hinge loss and the margin indicator loss function respectively. The case of the M-SVMs is dealt with in Section 6. At last, we draw conclusions and outline our ongoing research in Section 7.

2 Margin multi-category classifiers

The theoretical framework is that of [10]. It is summarized below.

2.1 Theoretical framework

We consider the case of C-category pattern classification problems with $C \in \mathbb{N} \setminus [\![0;2]\!]$. Each object is represented by its description $x \in \mathcal{X}$ and the set $\mathcal{Y}$ of the categories y can be identified with the set of indices of the categories: $[\![1;C]\!]$. We assume that $(\mathcal{X}, \mathcal{A}_\mathcal{X})$ and $(\mathcal{Y}, \mathcal{A}_\mathcal{Y})$ are measurable spaces and denote by $\mathcal{A}_\mathcal{X} \otimes \mathcal{A}_\mathcal{Y}$ the tensor-product sigma-algebra on the Cartesian product $\mathcal{X} \times \mathcal{Y}$. We make the hypothesis that the link between descriptions and categories can be characterized by an unknown probability measure P on the measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{A}_\mathcal{X} \otimes \mathcal{A}_\mathcal{Y})$. Let $Z = (X, Y)$ be a random pair with values in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, distributed according to P. The only access to P is via an m-sample $\mathbf{Z}_m = (Z_i)_{1 \leqslant i \leqslant m} = ((X_i, Y_i))_{1 \leqslant i \leqslant m}$ made up of independent copies of Z (in short $\mathbf{Z}_m \sim P^m$). The theoretical framework is thus that of *agnostic learning* [15]. The classifiers considered are based on classes of vector-valued functions with one component function per category, and the classes of component functions are *uniform Glivenko-Cantelli* (uGC) [9]. uGC classes must

be uniformly bounded up to additive constants (see Proposition 4 in [9]). For notational convenience, we replace this property by a stronger one: the vector-valued functions take their values in a hypercube of $\mathbb{R}^C$. The definition of a margin multi-category classifier is thus the following one.

Definition 1 (Margin multi-category classifiers) Let $\mathcal{G} = \prod_{k=1}^{C} \mathcal{G}_k$ be a class of functions from $\mathcal{X}$ into $[-M_{\mathcal{G}}, M_{\mathcal{G}}]^C$ with $M_{\mathcal{G}} \in [1, +\infty)$. The classes $\mathcal{G}_k$ of component functions are supposed to be uGC classes. For each function $g = (g_k)_{1 \leqslant k \leqslant C} \in \mathcal{G}$, a *margin multi-category classifier* on $\mathcal{X}$ is obtained by application of the *decision rule* dr, mapping g to $\mathrm{dr}_g \in (\mathcal{Y} \bigcup \{*\})^{\mathcal{X}}$, and defined as follows:

$$\forall x \in \mathcal{X}, \quad \begin{cases} \left|\mathrm{argmax}_{1 \leqslant k \leqslant C}\, g_k(x)\right| = 1 \Longrightarrow \mathrm{dr}_g(x) = \mathrm{argmax}_{1 \leqslant k \leqslant C}\, g_k(x) \\ \left|\mathrm{argmax}_{1 \leqslant k \leqslant C}\, g_k(x)\right| > 1 \Longrightarrow \mathrm{dr}_g(x) = * \end{cases}$$

where $|\cdot|$ returns the cardinality of its argument and $*$ stands for a dummy category.

In words, dr_g returns either the index of the component function whose value is the highest, or the dummy category $*$ in case of ex æquo. The qualifier *margin* refers to the fact that the generalization capabilities of such classifiers can be characterized by means of the values taken by the differences of the corresponding component functions. With this definition at hand, the aim of the *learning process* is to minimize over $\mathcal{G}$ the *probability of error* $P(\mathrm{dr}_g(X) \neq Y)$. This probability can be reformulated in a handy way thanks to the introduction of additional functions.

Definition 2 (Class of functions $\mathcal{F}_{\mathcal{G}}$) Let $\mathcal{G}$ be a class of functions satisfying Definition 1. For every $g \in \mathcal{G}$, the function f_g from $\mathcal{Z}$ into $[-M_{\mathcal{G}}, M_{\mathcal{G}}]$ is defined by:

$$\forall (x,k) \in \mathcal{Z}, \quad f_g(x,k) = \frac{1}{2}\left(g_k(x) - \max_{l \neq k} g_l(x)\right).$$

Then, the class $\mathcal{F}_{\mathcal{G}}$ is defined as follows: $\mathcal{F}_{\mathcal{G}} = \{f_g : \ g \in \mathcal{G}\}$.

Definition 3 (Risks) Let $\mathcal{G}$ be a class of functions satisfying Definition 1 and let ϕ be the standard indicator loss function given by: $\forall t \in \mathbb{R}, \quad \phi(t) = \mathbb{1}_{\{t \leqslant 0\}}$. The *expected risk* of any function $g \in \mathcal{G}$, $L(g)$, is given by: $L(g) = \mathbb{E}_{(X,Y)\sim P}[\phi \circ f_g(X,Y)] = P(\mathrm{dr}_g(X) \neq Y)$. Its *empirical risk* measured on the m-sample $\mathbf{Z}_m$ is: $L_m(g) = \mathbb{E}_{Z' \sim P_m}[\phi \circ f_g(Z')] = \frac{1}{m}\sum_{i=1}^{m} \phi \circ f_g(Z_i)$ (P_m is the empirical measure supported on $\mathbf{Z}_m$).

To benefit from the fact that the classifiers of interest are margin ones, the sample-based estimate of performance which is actually used is obtained by substituting to ϕ a (dominating) margin loss function parameterized by a scalar $\gamma \in (0,1]$: ϕ_γ. This gives birth to the margin risk $L_\gamma(g)$ and its empirical counterpart: $L_{\gamma,m}(g)$. In this study, the margin loss functions used are the following ones.

Definition 4 (Margin loss functions $\phi_{2,\gamma}$ and $\phi_{\infty,\gamma}$) For $\gamma \in (0,1]$, the *parameterized truncated hinge loss* $\phi_{2,\gamma}$ and the *margin indicator loss function* $\phi_{\infty,\gamma}$ are defined by

$$\forall t \in \mathbb{R}, \quad \begin{cases} \phi_{2,\gamma}(t) = \mathbb{1}_{\{t \leqslant 0\}} + \left(1 - \frac{t}{\gamma}\right) \mathbb{1}_{\{t \in (0,\gamma]\}} \\ \phi_{\infty,\gamma}(t) = \mathbb{1}_{\{t < \gamma\}} \end{cases} .$$

We use margin loss functions in combination with a squashing function.

Definition 5 (Piecewise-linear squashing function π_γ) For $\gamma \in (0,1]$, the function π_γ is defined by: $\forall t \in \mathbb{R}, \ \ \pi_\gamma(t) = t\mathbb{1}_{\{t \in (0,\gamma]\}} + \gamma \mathbb{1}_{\{t > \gamma\}}$.

The idea is to restrict the available information exactly to what is relevant for the assessment of the prediction accuracy, so as to optimize the exploitation of the scale parameter.

Definition 6 (Class of functions $\mathcal{F}_{\mathcal{G},\gamma}$) Let $\mathcal{G}$ be a class of functions satisfying Definition 1 and $\mathcal{F}_{\mathcal{G}}$ the class of functions deduced from $\mathcal{G}$ according to Definition 2. For every (ordered) pair $(g,\gamma) \in \mathcal{G} \times (0,1]$, the function $f_{g,\gamma}$ from $\mathcal{Z}$ into $[0,\gamma]$ is defined by: $f_{g,\gamma} = \pi_\gamma \circ f_g$. Then, the class $\mathcal{F}_{\mathcal{G},\gamma}$ is defined as follows: $\mathcal{F}_{\mathcal{G},\gamma} = \{f_{g,\gamma} : \ g \in \mathcal{G}\}$.

In the sequel, we make use of the floor function $\lfloor \cdot \rfloor$ and the ceiling function $\lceil \cdot \rceil$.

2.2 Scale-sensitive capacity measures

We introduce the three types of capacity measures, using the notations of [12]. Let $(\mathcal{T}, \mathcal{A}_{\mathcal{T}})$ be a measurable space and let $\mathcal{F}$ be a class of real-valued functions with domain $\mathcal{T}$. Let T be a random variable with values in $\mathcal{T}$, distributed according to a probability measure P_T on $(\mathcal{T}, \mathcal{A}_{\mathcal{T}})$ and let $\mathbf{T}_n = (T_i)_{1 \leqslant i \leqslant n}$ be an n-sample made up of independent copies of T. The empirical Rademacher complexity of $\mathcal{F}$ given $\mathbf{T}_n$ is denoted by $\hat{R}_n(\mathcal{F})$ and the Rademacher complexity of $\mathcal{F}$ is denoted by $R_n(\mathcal{F})$. The classes $\mathcal{F}$ considered here are endowed with empirical pseudo-metrics derived from the L_p-norm.

Definition 7 (Pseudo-distance $d_{p,\mathbf{t}_n}$) Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{T}$. For $n \in \mathbb{N}^*$, let $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$. Then,

$$\forall p \in [1,+\infty), \forall (f,f') \in \mathcal{F}^2, \ \ d_{p,\mathbf{t}_n}(f,f') = \|f - f'\|_{L_p(\mu_{\mathbf{t}_n})} = \left(\frac{1}{n} \sum_{i=1}^{n} |f(t_i) - f'(t_i)|^p \right)^{\frac{1}{p}}$$

and

$$\forall (f,f') \in \mathcal{F}^2, \ \ d_{\infty,\mathbf{t}_n}(f,f') = \|f - f'\|_{L_\infty(\mu_{\mathbf{t}_n})} = \max_{1 \leqslant i \leqslant n} |f(t_i) - f'(t_i)| ,$$

where $\mu_{\mathbf{t}_n}$ denotes the uniform (counting) probability measure on $\{t_i : 1 \leqslant i \leqslant n\}$.

Let $\bar{\mathcal{F}} \subset \mathcal{F}$. For $\epsilon \in \mathbb{R}_+^*$, $n \in \mathbb{N}^*$ and $p \in [1, +\infty]$, $\mathcal{N}\left(\epsilon, \bar{\mathcal{F}}, d_{p,\mathbf{t}_n}\right)$ and $\mathcal{M}\left(\epsilon, \bar{\mathcal{F}}, d_{p,\mathbf{t}_n}\right)$ respectively denote the ϵ-covering number and the ϵ-packing number of $\bar{\mathcal{F}}$ with respect to $d_{p,\mathbf{t}_n}$. $\mathcal{N}_p\left(\epsilon, \bar{\mathcal{F}}, n\right)$ and $\mathcal{M}_p\left(\epsilon, \bar{\mathcal{F}}, n\right)$ are the corresponding uniform covering and packing numbers. $\mathcal{N}^{\text{int}}$ and $\mathcal{N}_p^{\text{int}}$ are used to denote proper covering numbers. The multi-class scale-sensitive combinatorial dimension considered here is a γ-Ψ-dimension [10]: the margin Graph dimension.

Definition 8 (Graph dimension with margin γ) Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{Z}$. For $\gamma \in \mathbb{R}_+^*$, a subset $s_{\mathcal{Z}^n} = \{z_i : 1 \leqslant i \leqslant n\}$ of $\mathcal{Z}$ is said to be γ-*G-shattered* by $\mathcal{F}$ if there is a vector $\mathbf{b}_n = (b_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$ such that, for every vector $\mathbf{s}_n = (s_i)_{1 \leqslant i \leqslant n} \in \{-1, 1\}^n$, there is a function $f_{\mathbf{s}_n} \in \mathcal{F}$ satisfying

$$\forall i \in [\![1; n]\!], \quad \begin{cases} \text{if } s_i = 1, \ f_{\mathbf{s}_n}(x_i, y_i) - b_i \geqslant \gamma \\ \text{if } s_i = -1, \ \max_{k \neq y_i} f_{\mathbf{s}_n}(x_i, k) + b_i \geqslant \gamma \end{cases}.$$

The *Graph dimension with margin* γ of $\mathcal{F}$, denoted by γ-G-dim $(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{Z}$ γ-G-shattered by $\mathcal{F}$, if such maximum exists. Otherwise, $\mathcal{F}$ is said to have infinite Graph dimension with margin γ.

The γ-Ψ-dimensions generalize the standard scale-sensitive combinatorial dimension, named fat-shattering dimension or γ-dimension.

Definition 9 (γ-dimension [14]) Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{T}$. For $\gamma \in \mathbb{R}_+^*$, a subset $s_{\mathcal{T}^n} = \{t_i : 1 \leqslant i \leqslant n\}$ of $\mathcal{T}$ is said to be γ-*shattered* by $\mathcal{F}$ if there is a vector $\mathbf{b}_n = (b_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$ such that, for every vector $\mathbf{s}_n = (s_i)_{1 \leqslant i \leqslant n} \in \{-1, 1\}^n$, there is a function $f_{\mathbf{s}_n} \in \mathcal{F}$ satisfying

$$\forall i \in [\![1; n]\!], \ \ s_i \left(f_{\mathbf{s}_n}(t_i) - b_i\right) \geqslant \gamma.$$

The γ-dimension of the class $\mathcal{F}$, γ-dim $(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{T}$ γ-shattered by $\mathcal{F}$, if such maximum exists. Otherwise, $\mathcal{F}$ is said to have infinite γ-dimension.

The relationship between γ-G-dim $(\mathcal{F}_\mathcal{G})$ and γ-dim $(\mathcal{F}_\mathcal{G})$ is characterized by the following proposition.

Proposition 1 (After Proposition 1 in [13]) *Let $\mathcal{G}$ be a class of functions satisfying Definition 1 and $\mathcal{F}_\mathcal{G}$ the class of functions deduced from $\mathcal{G}$ according to Definition 2. Then,*

$$\forall \gamma \in (0, M_\mathcal{G}], \ \gamma\text{-}G\text{-}dim\left(\mathcal{F}_\mathcal{G}\right) \leqslant \gamma\text{-}dim\left(\mathcal{F}_\mathcal{G}\right).$$

The inequality becomes an equality for $C = 2$.

Each of the combinatorial results [28] in the literature is built upon a basic lemma involving classes of functions whose domain and codomain are finite sets (so that their cardinalities are also finite). To take benefit from this restriction, it involves a variant of the scale-sensitive combinatorial dimension considered. The first variant of this kind, associated with the γ-dimension, is the strong dimension [1], which extends to the margin Graph dimension as follows.

Definition 10 (Strong Graph dimension) Let $\mathcal{F}$ be a class of functions from $\mathcal{Z}$ into $[\![-M_{\mathcal{F}}; M_{\mathcal{F}}]\!]$ with $M_{\mathcal{F}} \in \mathbb{N}^*$. A subset $s_{\mathcal{Z}^n} = \{z_i : 1 \leqslant i \leqslant n\}$ of $\mathcal{Z}$ is said to be *strongly G-shattered* by $\mathcal{F}$ if there is a vector $\mathbf{b}_n = (b_i)_{1 \leqslant i \leqslant n} \in [\![-M_{\mathcal{F}} + 1; M_{\mathcal{F}} - 1]\!]$ such that, for every vector $\mathbf{s}_n = (s_i)_{1 \leqslant i \leqslant n} \in \{-1, 1\}^n$, there is a function $f_{\mathbf{s}_n} \in \mathcal{F}$ satisfying

$$\forall i \in [\![1; n]\!], \quad \begin{cases} \text{if } s_i = 1, \ f_{\mathbf{s}_n}(x_i, y_i) - b_i \geqslant 1 \\ \text{if } s_i = -1, \ \max_{k \neq y_i} f_{\mathbf{s}_n}(x_i, k) + b_i \geqslant 1 \end{cases} .$$

The *strong Graph dimension* of $\mathcal{F}$, denoted by S-G-dim $(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{Z}$ strongly G-shattered by $\mathcal{F}$, if such maximum exists. Otherwise, $\mathcal{F}$ is said to have infinite strong Graph dimension.

The finiteness of the domain results from a restriction to the data, whereas that of the codomain is obtained by application of a discretization operator. We make use of the one of [10].

Definition 11 (η-discretization operator) Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into the interval $[M_{\mathcal{F}-}, M_{\mathcal{F}+}]$. For $\eta \in \mathbb{R}_+^*$, define the *η-discretization* as an operator on $\mathcal{F}$ such that:

$$\begin{aligned} (\cdot)^{(\eta)} : \mathcal{F} &\longrightarrow \mathcal{F}^{(\eta)} \\ f &\mapsto f^{(\eta)} \end{aligned}$$

$$\forall t \in \mathcal{T}, \quad f^{(\eta)}(t) = \operatorname{sign}(f(t)) \cdot \left\lfloor \frac{|f(t)|}{\eta} \right\rfloor .$$

3 Connections between the capacity measures

The main building blocks of the derivation of guaranteed risks are connections between capacity measures. They are of two kinds. A first group corresponds to changes of capacity measure. A second group corresponds to structural results, basically relating the capacity of the multi-class classifier, precisely $\mathcal{F}_{\mathcal{G},\gamma}$ or $\mathcal{F}_{\mathcal{G}}$, to those of function classes including the classes $\mathcal{G}_k$.

3.1 Rademacher complexity

To the best of our knowledge, the sharpest structural result for classes of vector-valued functions is due to Maurer [20]. Under the simplifying hypotheses exposed in [13], its application to the framework of this study produces (up to a multiplicative factor), the following result, a direct consequence of the proof of Theorem 3 in [18].

Lemma 1 *Let $\mathcal{G}$ be a class of functions satisfying Definition 1. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 6. Then*

$$R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant CR_m\left(\bigcup_{k=1}^{C} \mathcal{G}_k\right).$$

Among the tools available to upper bound the expected suprema of empirical processes, an especially efficient one for Rademacher processes is Dudley's chaining method [25].

Theorem 1 (Dudley's metric entropy bound) *Let $\mathcal{F}$ be a class of bounded real-valued functions on $\mathcal{T}$. For $n \in \mathbb{N}^*$, let $\mathbf{t}_n \in \mathcal{T}^n$ and let $diam\left(\mathcal{F}\right) = \sup_{(f,f')\in\mathcal{F}^2} \left\|f - f'\right\|_{L_2(\mu_{\mathbf{t}_n})}$ be the diameter of $\mathcal{F}$ in the $L_2\left(\mu_{\mathbf{t}_n}\right)$ seminorm. Let h be a positive and decreasing function on $\mathbb{N}$ such that $h\left(0\right) \geqslant diam\left(\mathcal{F}\right)$. Then for $N \in \mathbb{N}^*$,*

$$\hat{R}_n\left(\mathcal{F}\right) \leqslant h\left(N\right) + 2\sum_{j=1}^{N}\left(h\left(j\right) + h\left(j-1\right)\right)\sqrt{\frac{\ln\left(\mathcal{N}^{int}\left(h\left(j\right),\mathcal{F},d_{2,\mathbf{t}_n}\right)\right)}{n}}.$$

3.2 Covering and packing numbers

The following structural result relates the covering numbers of $\mathcal{F}_{\mathcal{G},\gamma}$ to those of the classes $\mathcal{G}_k$.

Lemma 2 (Lemma 1 in [12]) *Let $\mathcal{G}$ be a class of functions satisfying Definition 1 and $\mathcal{F}_{\mathcal{G}}$ the class of functions deduced from $\mathcal{G}$ according to Definition 2. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 6. Then, for $\epsilon \in \mathbb{R}_+^*$, $n \in \mathbb{N}^*$, and $\mathbf{z}_n = \left(\left(x_i, y_i\right)\right)_{1\leqslant i\leqslant n} = \left(z_i\right)_{1\leqslant i\leqslant n} \in \mathcal{Z}^n$,*

$$\forall p \in [1,+\infty],\ \ \mathcal{N}^{int}\left(\epsilon,\mathcal{F}_{\mathcal{G},\gamma},d_{p,\mathbf{z}_n}\right) \leqslant \mathcal{N}^{int}\left(\epsilon,\mathcal{F}_{\mathcal{G}},d_{p,\mathbf{z}_n}\right) \leqslant \prod_{k=1}^{C}\mathcal{N}^{int}\left(\frac{\epsilon}{C^{\frac{1}{p}}},\mathcal{G}_k,d_{p,\mathbf{x}_n}\right),$$

where $\mathbf{x}_n = \left(x_i\right)_{1\leqslant i\leqslant n}$.

The transition between covering numbers and scale-sensitive combinatorial dimensions is obtained by application of a combinatorial result (generalized Sauer-Shelah lemma). This calls for the use of the following lemma.

Lemma 3 (After Theorem IV in [16]) *Let $(\mathcal{E}, \rho)$ be a pseudo-metric space. For every totally bounded set $\mathcal{E}' \subset \mathcal{E}$ and $\epsilon \in \mathbb{R}_+^*$, $\mathcal{N}^{int}(\epsilon, \mathcal{E}', \rho) \leqslant \mathcal{M}(\epsilon, \mathcal{E}', \rho)$.*

Restricting to the margin Graph dimension, one single generalized Sauer-Shelah lemma is available, that extends Theorem 1 in [22]. This dimension-free upper bound is based on the L_2-norm.

Lemma 4 (Lemma 7 in [13]) *Let $\mathcal{G}$ be a class of functions satisfying Definition 1 and $\mathcal{F}_\mathcal{G}$ the class of functions deduced from $\mathcal{G}$ according to Definition 2. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 6. For $\epsilon \in (0, M_\mathcal{G}]$, let $d_G(\epsilon) = \epsilon\text{-}G\text{-}dim(\mathcal{F}_\mathcal{G})$. Then for $\epsilon \in (0,\gamma]$ and $n \in \mathbb{N}^*$,*

$$\mathcal{M}_2(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, n) \leqslant \left(\frac{6\gamma}{\epsilon}\right)^{20 d_G\left(\frac{\epsilon}{48}\right)}.$$

We now derive an extension to the margin Graph dimension of the best generalized Sauer-Shelah lemma based on the L_∞-norm: Lemma 3.5 in [1]. To that end, two main technical lemmas are to be introduced. The first one realizes the transition to and from the discretized setting.

Lemma 5 (After Lemma 3.2 in [1]) *Let $\mathcal{G}$ be a class of functions satisfying Definition 1 and $\mathcal{F}_\mathcal{G}$ the class of functions deduced from $\mathcal{G}$ according to Definition 2. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 6. For $n \in \mathbb{N}^*$, let $\mathbf{z}_n \in \mathcal{Z}^n$.*

1. *For every $\epsilon \in (0,\gamma]$ and every $\eta \in \left(0, \frac{\epsilon}{2}\right]$,*

$$\mathcal{M}(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, d_{\infty,\mathbf{z}_n}) \leqslant \mathcal{M}\left(2, \mathcal{F}_{\mathcal{G},\gamma}^{(\eta)}, d_{\infty,\mathbf{z}_n}\right);$$

2. *for every $\eta \in (0, M_\mathcal{G}]$ and every $\epsilon \in \left(0, \frac{\eta}{2}\right]$,*

$$S\text{-}G\text{-}dim\left(\mathcal{F}_\mathcal{G}^{(\eta)}\right) \leqslant \epsilon\text{-}G\text{-}dim(\mathcal{F}_\mathcal{G}).$$

The following lemma, connecting separation and strong G-shattering, is at the core of the basic combinatorial result.

Lemma 6 (After Lemma 13 in [13]) *Let $\mathcal{G}$ be a class of functions satisfying Definition 1 and $\mathcal{F}_\mathcal{G}$ the class of functions deduced from $\mathcal{G}$ according to Definition 2. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 6. Let $\eta \in \left(0, \frac{\gamma}{2}\right]$. Suppose that there exist $(g, g') \in \mathcal{G}^2$ and $z \in \mathcal{Z}$ such that*

$$f_{g,\gamma}^{(\eta)}(z) - f_{g',\gamma}^{(\eta)}(z) \geqslant 2.$$

Then the set $\left\{f_g^{(\eta)}, f_{g'}^{(\eta)}\right\}$ strongly G-shatters the singleton $\{z\}$. Furthermore, a witness to this shattering is $b = f_{g,\gamma}^{(\eta)}(z) - 1$.

The second main technical lemma is the basic combinatorial result. Its formulation calls for the introduction of additional notations. For $\gamma \in (0,1]$, $\eta \in \left(0, \frac{\gamma}{2}\right]$ and $s_{\mathcal{Z}^n} = \{z_i : 1 \leqslant i \leqslant n\} \subset \mathcal{Z}$, let $\mathcal{F}_\gamma$ be a subset of $\left(\mathcal{F}_{\mathcal{G},\gamma}|_{s_{\mathcal{Z}^n}}\right)^{(\eta)}$ and let $\mathcal{F}$ be a subset of $\left(\mathcal{F}_{\mathcal{G}}|_{s_{\mathcal{Z}^n}}\right)^{(\eta)}$. $\mathcal{F}$ will be said to correspond to $\mathcal{F}_\gamma$ if there exists a subset $\tilde{\mathcal{G}}$ of $\mathcal{G}$ of cardinality $|\mathcal{F}_\gamma|$ such that $\mathcal{F}_\gamma = \left\{\left(f_{g,\gamma}|_{s_{\mathcal{Z}^n}}\right)^{(\eta)} : g \in \tilde{\mathcal{G}}\right\}$ and $\mathcal{F} = \left\{\left(f_g|_{s_{\mathcal{Z}^n}}\right)^{(\eta)} : g \in \tilde{\mathcal{G}}\right\}$. Given a vector $\mathbf{z}_n = (z_i)_{1\leqslant i\leqslant n} \in \mathcal{Z}^n$, $s_{\mathcal{Z}^{n'}}$ will designate the smallest subset of $\mathcal{Z}$ including all the components of $\mathbf{z}_n$ ($n' \leqslant n$).

Lemma 7 (After Lemma 3.3 in [1]) *Let $\mathcal{G}$ be a class of functions satisfying Definition 1 and $\mathcal{F}_\mathcal{G}$ the class of functions deduced from $\mathcal{G}$ according to Definition 2. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 6. For $\mathbf{z}_n = (z_i)_{1\leqslant i\leqslant n} \in \mathcal{Z}^n$, $\gamma \in (0,1]$ and $\eta \in \left(0, \frac{\gamma}{2}\right]$, let $\mathcal{F}_\gamma$ be a subset of $\left(\mathcal{F}_{\mathcal{G},\gamma}|_{s_{\mathcal{Z}^{n'}}}\right)^{(\eta)}$ and let $\mathcal{F}$ be a subset of $\left(\mathcal{F}_{\mathcal{G}}|_{s_{\mathcal{Z}^{n'}}}\right)^{(\eta)}$ corresponding to $\mathcal{F}_\gamma$. Suppose that $\mathcal{M}\left(2, \mathcal{F}_\gamma, d_{\infty,\mathbf{z}_n}\right) \geqslant 2$. Then,*

$$\mathcal{M}\left(2, \mathcal{F}_\gamma, d_{\infty,\mathbf{z}_n}\right) < 2\left(\left\lfloor \frac{\gamma}{\eta} \right\rfloor^2 n\right)^{\lceil \log_2(\Sigma)\rceil},$$

where $\Sigma = \sum_{u=1}^{d} \binom{n}{u} \left(\left\lfloor \frac{\gamma}{\eta} \right\rfloor - 1\right)^u$ with $d = \text{S-G-dim}\left(\mathcal{F}\right)$.

Proof Notice first that according to Lemma 6,

$$\mathcal{M}\left(2, \mathcal{F}_\gamma, d_{\infty,\mathbf{z}_n}\right) \geqslant 2 \Longrightarrow d \geqslant 1.$$

For notational convenience, we focus on the case leading to the largest bound: that corresponding to $n' = n$ (all the components of $\mathbf{z}_n$ are distinct elements of $\mathcal{Z}$). Furthermore, we perform implicit permutations on sets of indexed elements when these permutations are obvious. We also set $M_\gamma = \left\lfloor \frac{\gamma}{\eta} \right\rfloor$. For $u \in [\![1; d]\!]$, let us say that the class $\bar{\mathcal{F}} \subset \mathcal{F}$ strongly G-shatters a pair $(s_{\mathcal{Z}^u}, \mathbf{b}_u)$, where $s_{\mathcal{Z}^u}$ is a subset of $s_{\mathcal{Z}^n}$ of cardinality u and $\mathbf{b}_u \in [\![1; M_\gamma - 1]\!]^u$, if $\bar{\mathcal{F}}$ strongly G-shatters $s_{\mathcal{Z}^u}$ according to $\mathbf{b}_u$. Note that since $\bar{\mathcal{F}} \subset \mathcal{F}$, S-G-dim $\left(\bar{\mathcal{F}}\right) \leqslant d$ so that the number of such pairs is inferior or equal to

$$\Sigma = \sum_{u=1}^{d} \binom{n}{u} \left(M_\gamma - 1\right)^u.$$

For $q \in [\![2; |\mathcal{F}_\gamma|]\!]$ and $r \in [\![1; n]\!]$, let us define the set $\mathcal{S}\left(q, r\right)$ of subsets of $\mathcal{F}_\gamma$ as follows:

$$\bar{\mathcal{F}}_\gamma \in \mathcal{S}\left(q, r\right) \Longleftrightarrow \begin{cases} \left|\bar{\mathcal{F}}_\gamma\right| = q \\ \exists \mathbf{z}_r \in \mathcal{Z}^r : \forall \left\{f_\gamma, f'_\gamma\right\} \subset \bar{\mathcal{F}}_\gamma,\ d_{\infty,\mathbf{z}_r}\left(f_\gamma, f'_\gamma\right) \geqslant 2 \end{cases},$$

where $\mathbf{z}_r$ is a subvector of $\mathbf{z}_n$ of size r. Then, $h(q,r)$ denotes the maximum number such that, for every subset $\bar{\mathcal{F}}_\gamma$ of $\mathcal{F}_\gamma$ belonging to $\mathcal{S}(q,r)$, the corresponding subset of $\mathcal{F}$ stronly G-shatters at least that many pairs. If $\mathcal{S}(q,r) = \emptyset$, then $h(q,r)$ is infinite. Given the definitions of Σ and the function h,

$$h(q,n) > \Sigma \Longrightarrow \mathcal{M}(2, \mathcal{F}_\gamma, d_{\infty,\mathbf{z}_n}) < q.$$

Therefore, to finish the proof, it suffices to show that

$$h\left(2\left(M_\gamma^2 n\right)^{\lceil \log_2(\Sigma)\rceil}, n\right) > \Sigma. \tag{1}$$

To that end, two cases are considered.

• $\lceil \log_2(\Sigma)\rceil \geqslant n$. Then, $2\left(M_\gamma^2 n\right)^{\lceil \log_2(\Sigma)\rceil}$ is strictly larger than the total number of functions from $s_{\mathcal{Z}^n}$ to $[\![0; M_\gamma]\!]$, $(M_\gamma+1)^n$, and thus also strictly larger than $|\mathcal{F}_\gamma|$. Consequently, by definition of the function h, $h\left(2\left(M_\gamma^2 n\right)^{\lceil \log_2(\Sigma)\rceil}, n\right) = +\infty > \Sigma$. We now turn to the second case.

• $\lceil \log_2(\Sigma)\rceil < n$. In that second case, the core of the proof is the proof of the following claim:

$$\begin{cases} \forall r \in [\![1; n]\!],\ h(2,r) \geqslant 1 \\ \forall r \in [\![2; n]\!], \forall K \in \mathbb{N}^*,\ h\left(2KM_\gamma^2 r, r\right) > 2h(2K, r-1) \end{cases}.$$

The first part of the claim is a direct consequence of Lemma 6. For the second part, first note that if $\mathcal{S}\left(2KM_\gamma^2 r, r\right) = \emptyset$, then $h\left(2KM_\gamma^2 r, r\right) = +\infty$ and hence the claim holds. To complete the proof of the claim, we can thus assume that there exist a subvector $\mathbf{z}_r$ of $\mathbf{z}_n$ and $\mathcal{F}_{\gamma,2} \subset \mathcal{F}_\gamma$ whose $2KM_\gamma^2 r$ functions are pairwise 2-separated with respect to $d_{\infty,\mathbf{z}_r}$. Split $\mathcal{F}_{\gamma,2}$ arbitrarily into $KM_\gamma^2 r$ pairs. For each pair $\left(f_\gamma, f'_\gamma\right)$, find $z_i \in s_{\mathcal{Z}^r}$ such that $\left|f_\gamma(z_i) - f'_\gamma(z_i)\right| \geqslant 2$. By the pigeonhole principle, the same example is picked for at least KM_γ^2 pairs. Let z_{i_0} be such an example. Notice that the number of pairs $(k_+, k_-) \in [\![0; M_\gamma]\!]^2$ satisfying $k_+ \geqslant k_- + 2$ is equal to $\frac{M_\gamma(M_\gamma-1)}{2}$. Consequently, a second application of the pigeonhole principle establishes that there are at least $\left\lceil \frac{2KM_\gamma^2}{M_\gamma(M_\gamma-1)} \right\rceil \geqslant 2K$ of the pairs for which the (unordered) pair $\left(f_\gamma(z_{i_0}), f'_\gamma(z_{i_0})\right)$ is the same. For all the pairs, let us transpose the functions if needed so that we have systematically: $f_\gamma(z_{i_0}) - f'_\gamma(z_{i_0}) \geqslant 2$. The corresponding sets $\{f, f'\} \subset \mathcal{F}$ all shatter $\{z_{i_0}\}$ (shatter at least one pair of the form $(\{z_{i_0}\}, b_{i_0})$). Furthermore, according to Lemma 6, for each of these sets, b_{i_0} can be set equal to $f_\gamma(z_{i_0}) - 1$ (which belongs to $[\![1; M_\gamma - 1]\!]$ as required). To sum up, there are two subsets of $\mathcal{F}_{\gamma,2}$, call them $\mathcal{F}_{\gamma,+}$ and $\mathcal{F}_{\gamma,-}$ and there are $z_{i_0} \in s_{\mathcal{Z}^r}$ and $(k_+, k_-) \in [\![0; M_\gamma]\!]^2$ with $k_+ \geqslant k_- + 2$ so that

$|\mathcal{F}_{\gamma,+}| = |\mathcal{F}_{\gamma,-}| = 2K$, for every $f_{\gamma,+} \in \mathcal{F}_{\gamma,+}$, $f_{\gamma,+}(z_{i_0}) = k_+$, for every $f_{\gamma,-} \in \mathcal{F}_{\gamma,-}$, $f_{\gamma,-}(z_{i_0}) = k_-$, and

$$\forall (f_{\gamma,+}, f_{\gamma,-}) \in \mathcal{F}_{\gamma,+} \times \mathcal{F}_{\gamma,-}, \quad \begin{cases} f_+(x_{i_0}, y_{i_0}) - b_{i_0} \geqslant 1 \\ \max_{k \neq y_{i_0}} f_-(x_{i_0}, k) + b_{i_0} \geqslant 1 \end{cases},$$

where $\{f_+, f_-\}$ is the subset of $\mathcal{F}$ corresponding to $\{f_{\gamma,+}, f_{\gamma,-}\}$. Since the functions in $\mathcal{F}_{\gamma,+}$ are pairwise 2-separated with respect to $d_{\infty,\mathbf{z}_r}$ but take the same value on z_{i_0}, they are also pairwise 2-separated with respect to $d_{\infty,\mathbf{z}_{r-1}}$ (with $\mathbf{z}_{r-1}$ being the vector associated with the set $s_{\mathcal{Z}^{r-1}} = s_{\mathcal{Z}^r} \setminus \{z_{i_0}\}$). The same holds for the functions in $\mathcal{F}_{\gamma,-}$. Hence, both $\mathcal{F}_{\gamma,+}$ and $\mathcal{F}_{\gamma,-}$ belong to $\mathcal{S}(2K, r-1)$. By the definition of the function h, the corresponding subsets of $\mathcal{F}$, respectively $\mathcal{F}_+$ and $\mathcal{F}_-$, strongly G-shatter at least $h(2K, r-1)$ pairs $(s_{\mathcal{Z}^u}, \mathbf{b}_u)$. Obviously, $\mathcal{F}_2$, the subset of $\mathcal{F}$ corresponding to $\mathcal{F}_{\gamma,2}$, strongly G-shatters all the pairs strongly G-shattered by either $\mathcal{F}_+$ or $\mathcal{F}_-$ plus $(\{z_{i_0}\}, b_{i_0})$. Moreover, if the same pair $(s_{\mathcal{Z}^u}, \mathbf{b}_u)$ is strongly G-shattered by both $\mathcal{F}_+$ and $\mathcal{F}_-$, then $\mathcal{F}_2$ also strongly G-shatters the pair $\left(s'_{\mathcal{Z}^{u+1}}, \mathbf{b}'_{u+1}\right)$, where $s'_{\mathcal{Z}^{u+1}}$ is the set deduced from $s_{\mathcal{Z}^u}$ by adding one point $z'_{u+1} = z_{i_0}$ and the vector $\mathbf{b}'_{u+1}$ is deduced from $\mathbf{b}_u$ by appending one component $b'_{u+1} = b_{i_0}$. Clearly, neither $\mathcal{F}_+$ nor $\mathcal{F}_-$ strongly G-shatters this pair, simply because they do not strongly G-shatter the pair $(\{z_{i_0}\}, b_{i_0})$. It follows that $\mathcal{F}_2$ strongly G-shatters at least $2h(2K, r-1) + 1$ pairs, from which it stems that $h\left(2KM_\gamma^2 r, r\right) > 2h(2K, r-1)$, completing the proof of the claim.

Now for $u \in [\![1; n-1]\!]$, let $q = 2\left(M_\gamma^2\right)^u \Pi_{j=1}^u (n-j+1)$. By repeated application of the above claim, it follows that $h(q, n) > 2^u$. Since the function h is clearly nondecreasing in its first argument and $2\left(M_\gamma^2 n\right)^u \geqslant q$, this implies that

$$\forall u \in [\![1; n-1]\!],\ h\left(2\left(M_\gamma^2 n\right)^u, n\right) > 2^u. \tag{2}$$

Since the case considered is $\lceil \log_2(\Sigma) \rceil < n$, Inequality (1) then results from setting $u = \lceil \log_2(\Sigma) \rceil$ in Inequality (2). Indeed, it yealds to

$$\begin{aligned} h\left(2\left(M_\gamma^2 n\right)^{\lceil \log_2(\Sigma) \rceil}, n\right) &> 2^{\lceil \log_2(\Sigma) \rceil} \\ &\geqslant \Sigma. \end{aligned}$$

Thus, Inequality (1) holds true in both cases ($\lceil \log_2(\Sigma) \rceil \geqslant n$ and $\lceil \log_2(\Sigma) \rceil < n$), which completes the proof of the lemma.

With Lemmas 5 and 7 at hand, we obtain the extension of Lemma 3.5 in [1] announced.

Lemma 8 *Let $\mathcal{G}$ be a class of functions satisfying Definition 1 and $\mathcal{F}_\mathcal{G}$ the class of functions deduced from $\mathcal{G}$ according to Definition 2. For $\gamma \in (0, 1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 6.*

For $\epsilon \in (0, M_{\mathcal{G}}]$, let $d_G(\epsilon) = \epsilon$-G-dim$(\mathcal{F}_{\mathcal{G}})$. Then for $\gamma \in (0,1]$, $\epsilon \in (0,\gamma]$ and $n \in \mathbb{N}^*$ such that $n \geqslant d_G\left(\frac{\epsilon}{4}\right)$,

$$\mathcal{M}_\infty(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, n) < 2\left(\frac{4\gamma^2 n}{\epsilon^2}\right)^{\left\lceil d_G\left(\frac{\epsilon}{4}\right)\log_2\left(\frac{2\gamma e n}{d_G\left(\frac{\epsilon}{4}\right)\epsilon}\right)\right\rceil}. \tag{3}$$

Proof First, note that Inequality (3) is trivially true for $\mathcal{M}_\infty(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, n) < 2$. Thus, we proceed under the complementary hypothesis. Let $\mathbf{z}_n = (z_i)_{1 \leqslant i \leqslant n} \in \mathcal{Z}^n$ be such that $\mathcal{M}(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, d_{\infty,\mathbf{z}_n}) \geqslant 2$. Without loss of generality, we can assume that all the components of this vector are distinct elements of $\mathcal{Z}$ and thus define accordingly $s_{\mathcal{Z}^n} = \{z_i : 1 \leqslant i \leqslant n\} \subset \mathcal{Z}$. Indeed, this is the case leading to the largest bound. By definition,

$$\mathcal{M}(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, d_{\infty,\mathbf{z}_n}) = \mathcal{M}\left(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}|_{s_{\mathcal{Z}^n}}, d_{\infty,\mathbf{z}_n}\right).$$

Furthermore, setting $\eta = \frac{\epsilon}{2}$ in the first proposition of Lemma 5, one obtains:

$$\mathcal{M}\left(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}|_{s_{\mathcal{Z}^n}}, d_{\infty,\mathbf{z}_n}\right) \leqslant \mathcal{M}\left(2, \left(\mathcal{F}_{\mathcal{G},\gamma}|_{s_{\mathcal{Z}^n}}\right)^{\left(\frac{\epsilon}{2}\right)}, d_{\infty,\mathbf{z}_n}\right).$$

Since by hypothesis, $\mathcal{M}\left(2, \left(\mathcal{F}_{\mathcal{G},\gamma}|_{s_{\mathcal{Z}^n}}\right)^{\left(\frac{\epsilon}{2}\right)}, d_{\infty,\mathbf{z}_n}\right) \geqslant 2$, the packing number of $\left(\mathcal{F}_{\mathcal{G},\gamma}|_{s_{\mathcal{Z}^n}}\right)^{\left(\frac{\epsilon}{2}\right)}$ can be upper bounded thanks to Lemma 7, leading to

$$\mathcal{M}\left(2, \left(\mathcal{F}_{\mathcal{G},\gamma}|_{s_{\mathcal{Z}^n}}\right)^{\left(\frac{\epsilon}{2}\right)}, d_{\infty,\mathbf{z}_n}\right) < 2\left(\left\lfloor\frac{2\gamma}{\epsilon}\right\rfloor^2 n\right)^{\lceil\log_2(\Sigma)\rceil},$$

where $\Sigma = \sum_{u=1}^{d}\binom{n}{u}\left(\left\lfloor\frac{2\gamma}{\epsilon}\right\rfloor - 1\right)^u$ with $d = \text{S-G-dim}\left(\left(\mathcal{F}_{\mathcal{G}}|_{s_{\mathcal{Z}^n}}\right)^{\left(\frac{\epsilon}{2}\right)}\right)$. Making use of the second proposition of Lemma 5 gives:

$$\begin{aligned}
\text{S-G-dim}\left(\left(\mathcal{F}_{\mathcal{G}}|_{s_{\mathcal{Z}^n}}\right)^{\left(\frac{\epsilon}{2}\right)}\right) &\leqslant \frac{\epsilon}{4}\text{-G-dim}\left(\mathcal{F}_{\mathcal{G}}|_{s_{\mathcal{Z}^n}}\right) \\
&\leqslant \frac{\epsilon}{4}\text{-G-dim}\left(\mathcal{F}_{\mathcal{G}}\right) \\
&= d_G\left(\frac{\epsilon}{4}\right).
\end{aligned}$$

As a consequence, since by hypothesis, $n \geqslant d_G\left(\frac{\epsilon}{4}\right)$, a well-known computation (see for instance the proof of Corollary 3.3 in [23]) provides

$$\begin{aligned}\Sigma &= \sum_{u=1}^{d}\binom{n}{u}\left(\left\lfloor\frac{2\gamma}{\epsilon}\right\rfloor-1\right)^u \\ &\leqslant \left(\frac{2\gamma}{\epsilon}\right)^{d_G\left(\frac{\epsilon}{4}\right)}\sum_{u=1}^{d_G\left(\frac{\epsilon}{4}\right)}\binom{n}{u} \\ &\leqslant \left(\frac{2\gamma e n}{d_G\left(\frac{\epsilon}{4}\right)\epsilon}\right)^{d_G\left(\frac{\epsilon}{4}\right)}.\end{aligned}$$

Putting things together, so far, we have established that

$$\mathcal{M}\left(\epsilon,\mathcal{F}_{\mathcal{G},\gamma},d_{\infty,\mathbf{z}_n}\right) < 2\left(\frac{4\gamma^2 n}{\epsilon^2}\right)^{\left\lceil d_G\left(\frac{\epsilon}{4}\right)\log_2\left(\frac{2\gamma e n}{d_G\left(\frac{\epsilon}{4}\right)\epsilon}\right)\right\rceil}.$$

To complete the proof of (3), it suffices to notice that the right-hand side does not depend on $\mathbf{z}_n$ (but only on n).

It is noteworthy that the proofs of Lemmas 4 and 8 can be extended in a straightforward way so as to obtain bounds where the margin Graph dimension of $\mathcal{F}_{\mathcal{G}}$ is replaced with its fat-shattering dimension. However, this comes without any improvement, so that according to Proposition 1, the bounds are simply worsened. This observation backs the thesis that guaranteed risks for margin multi-category classifiers should be based on γ-Ψ-dimensions of $\mathcal{F}_{\mathcal{G}}$ rather than on its fat-shattering dimension.

3.3 Scale-sensitive combinatorial dimensions

The best decomposition result available for ϵ-G-dim $(\mathcal{F}_{\mathcal{G}})$ is Lemma 8 in [13]. This general purpose structural result can be significantly improved for specific classifiers such as the M-SVMs [11,19,8]. To take this gap into account, we resort to the following hypothesis, built upon an hypothesis which is standard in learning theory [21,12], and beyond in the theory of empirical processes [26]: that of polynomial γ-dimensions.

Hypothesis 1 *We consider classes of functions $\mathcal{G}$ satisfying Definition 1 plus the fact that there exists a quadruplet $(d_{\mathcal{G},C}, d_{\mathcal{G},\gamma}, K_{\mathcal{G},1}, K_{\mathcal{G},2}) \in \left(\mathbb{R}_+^*\right)^4$ such that for every $\epsilon \in (0, M_{\mathcal{G}}]$,*

$$\epsilon\text{-}G\text{-}dim\left(\mathcal{F}_{\mathcal{G}}\right) \leqslant K_{\mathcal{G},1}C^{d_{\mathcal{G},C}}\max_{1\leqslant k\leqslant C}\epsilon\text{-}dim\left(\mathcal{G}_k\right) \leqslant K_{\mathcal{G},2}C^{d_{\mathcal{G},C}}\epsilon^{-d_{\mathcal{G},\gamma}}.$$

Regarding the polynomial growth of the γ-dimensions, Corollary 27 in [3] tells us that if $\mathcal{F}$ is the class of functions computed by a bi-class multi-layer perceptron (MLP) [2], then there exists $K_{\mathcal{F}} \in \mathbb{R}_+^*$ such that $\epsilon\text{-dim}\,(\mathcal{F}) \leqslant K_{\mathcal{F}} \epsilon^{-(2n+2)}$ where n is the number of hidden layers. According to Theorem 4.6 in [4], the same result holds true, with an exponent equal to -2, if $\mathcal{F}$ is the class of functions computed by a support vector machine (SVM) [6]. Since our study focuses on the dependence on the three basic parameters rather than on the values of the constants, in the sequel, to simplify notations, $K_{\mathcal{G},1}$ and $K_{\mathcal{G},2}$ are replaced with one single constant: $K_{\mathcal{G}}$.

4 New guaranteed risks based on the truncated hinge loss

For this loss function as for the indicator loss function (see Section 5), we start with a basic supremum inequality from the literature and upper bound its capacity measure in such a way as to obtain an explicit dependence on m, C and γ.

4.1 Basic supremum inequality

The basic supremum inequality involving $\phi_{2,\gamma}$ appears (for instance) as a partial result in the proof of Theorem 8.1 in [23] (with $\mathcal{F}_{\mathcal{G}}$ replaced with $\mathcal{F}_{\mathcal{G},\gamma}$).

Theorem 2 *Let $\mathcal{G}$ be a class of functions satisfying Definition 1. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 6. For a fixed $\gamma \in (0,1]$ and a fixed $\delta \in (0,1)$, with P^m-probability at least $1-\delta$, uniformly for every function $g \in \mathcal{G}$,*

$$L\,(g) \leqslant L_{\gamma,m}\,(g) + \frac{2}{\gamma} R_m\,(\mathcal{F}_{\mathcal{G},\gamma}) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}},$$

where the margin loss function is $\phi_{2,\gamma}$.

4.2 Upper bounds on the Rademacher complexity

To upper bound the Rademacher complexity of interest, we consider the three options corresponding to the different levels at which the decomposition from the multi-class case to the bi-class case can be implemented (see Section 3).

4.2.1 Decomposition involving Rademacher complexities

In this case, to handle more easily the term $\bigcup_{k=1}^{C} \mathcal{G}_k$ that Lemma 1 makes appear, we add an hypothesis widely verified in practice: $\mathcal{G} = \mathcal{G}_0^C$ (all the classes of component functions are identical, with the consequence that $\bigcup_{k=1}^{C} \mathcal{G}_k = \mathcal{G}_0$). Examples of such classifiers are the MLPs and the M-SVMs. Applying in

sequence Lemma 1, Theorem 1, Lemma 3, Theorem 1 in [22] and Hypothesis 1, and setting $d(\epsilon) = \epsilon\text{-dim}(\mathcal{G}_0)$ gives:

$$\begin{aligned} R_m(\mathcal{F}_{\mathcal{G},\gamma}) &\leqslant CR_m(\mathcal{G}_0) \\ &\leqslant C\left(h(N) + 2\sum_{j=1}^{N}(h(j)+h(j-1))\sqrt{\frac{\ln\left(\mathcal{N}_2^{\text{int}}(h(j),\mathcal{G}_0,m)\right)}{m}}\right) \\ &\leqslant C\left(h(N) + 4\sqrt{\frac{5}{m}}\sum_{j\in\mathcal{J}}(h(j)+h(j-1))\sqrt{d\left(\frac{h(j)}{48}\right)\ln\left(\frac{12M_{\mathcal{G}}}{h(j)}\right)}\right) \\ &\leqslant C\left(h(N) + 4\sqrt{\frac{5\cdot 48^{d_{\mathcal{G},\gamma}}K_{\mathcal{G}}}{m}}\sum_{j\in\mathcal{J}}\frac{h(j)+h(j-1)}{h(j)^{\frac{d_{\mathcal{G},\gamma}}{2}}}\sqrt{\ln\left(\frac{12M_{\mathcal{G}}}{h(j)}\right)}\right) \end{aligned}$$

where $\mathcal{J} = \{j \in [\![1;N]\!] : h(j) \leqslant 2M_{\mathcal{G}}\}$. As in [12], the set $\mathcal{J}$ has been introduced to take into account the hypotheses of the generalized Sauer-Shelah lemma (here Theorem 1 in [22]) This raises no difficulties since $h(j) > 2M_{\mathcal{G}} \Longrightarrow \ln\left(\mathcal{N}_2^{\text{int}}(h(j),\mathcal{G}_0,m)\right) = 0$.

Theorem 3 *Let $\mathcal{G}$ be a class of functions satisfying Hypothesis 1. Suppose further that there exists a function class $\mathcal{G}_0$ such that $\mathcal{G} = \mathcal{G}_0^C$. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 6.*
If $d_{\mathcal{G},\gamma} \in (0,2)$, then

$$R_m(\mathcal{F}_{\mathcal{G},\gamma}) \leqslant 8\frac{1+2^{\frac{2}{2-d_{\mathcal{G},\gamma}}}}{\sqrt{2(2-d_{\mathcal{G},\gamma})}}(2M_{\mathcal{G}})^{1-\frac{d_{\mathcal{G},\gamma}}{2}}\sqrt{\frac{5\cdot 48^{d_{\mathcal{G},\gamma}}K_{\mathcal{G}}}{m}}C\left[\sqrt{\ln(F(6))}+\sqrt{\frac{1}{4\ln(F(6))}}\right] \tag{4}$$

where $F(t) = 2\cdot t^{\frac{2-d_{\mathcal{G},\gamma}}{2}}$.
If $d_{\mathcal{G},\gamma} = 2$, then

$$R_m(\mathcal{F}_{\mathcal{G},\gamma}) \leqslant \frac{2M_{\mathcal{G}}C}{\sqrt{m}}\left(1++576\sqrt{5K_{\mathcal{G}}}(2M_{\mathcal{G}})^{-1}\left\lceil\frac{1}{2}\log_2(m)\right\rceil\sqrt{\ln\left(6\sqrt{m}\right)}\right).$$

At last, if $d_{\mathcal{G},\gamma} > 2$, then

$$R_m(\mathcal{F}_{\mathcal{G},\gamma}) \leqslant 2M_{\mathcal{G}}C\left(\frac{\log_2(m)}{m}\right)^{\frac{1}{d_{\mathcal{G},\gamma}}}\left(1+8\left(1+2^{\frac{2}{d_{\mathcal{G},\gamma}-2}}\right)(2M_{\mathcal{G}})^{-\frac{d_{\mathcal{G},\gamma}}{2}}\sqrt{\frac{5\cdot 48^{d_{\mathcal{G},\gamma}}K_{\mathcal{G}}}{\log_2(m)}}\sqrt{\ln\left(6\left(\frac{m}{\log_2(m)}\right)^{\frac{1}{d_{\mathcal{G},\gamma}}}\right)}\right).$$

Proof
First case: $d_{\mathcal{G},\gamma} \in (0,2)$
This case is the only one for which the entropy integral exists. Setting for every

$j \in \mathbb{N}$, $h(j) = 2M_{\mathcal{G}} 2^{-\frac{2}{2-d_{\mathcal{G},\gamma}} j}$, we obtain

$$\begin{aligned} R_m(\mathcal{F}_{\mathcal{G},\gamma}) &\leqslant 8\left(1+2^{\frac{2}{2-d_{\mathcal{G},\gamma}}}\right)(2M_{\mathcal{G}})^{1-\frac{d_{\mathcal{G},\gamma}}{2}}\sqrt{\frac{5\cdot 48^{d_{\mathcal{G},\gamma}}K_{\mathcal{G}}}{m}}C\int_0^{\frac{1}{2}}\sqrt{\ln\left(\frac{6}{\epsilon^{\frac{2}{2-d_{\mathcal{G},\gamma}}}}\right)}d\epsilon \\ &= 8\frac{\left(1+2^{\frac{2}{2-d_{\mathcal{G},\gamma}}}\right)\sqrt{2}}{\sqrt{2-d_{\mathcal{G},\gamma}}}(2M_{\mathcal{G}})^{1-\frac{d_{\mathcal{G},\gamma}}{2}}\sqrt{\frac{5\cdot 48^{d_{\mathcal{G},\gamma}}K_{\mathcal{G}}}{m}}C\int_0^{\frac{1}{2}}\sqrt{\ln\left(\frac{6^{\frac{2-d_{\mathcal{G},\gamma}}{2}}}{\epsilon}\right)}d\epsilon. \end{aligned}$$

Let us define the integral $I(t)$ as follows:

$$I(t) = \int_0^{\frac{1}{2}}\sqrt{\ln\left(\frac{F(t)}{2\epsilon}\right)}d\epsilon.$$

Then,

$$R_m(\mathcal{F}_{\mathcal{G},\gamma}) \leqslant 8\frac{\left(1+2^{\frac{2}{2-d_{\mathcal{G},\gamma}}}\right)\sqrt{2}}{\sqrt{2-d_{\mathcal{G},\gamma}}}(2M_{\mathcal{G}})^{1-\frac{d_{\mathcal{G},\gamma}}{2}}\sqrt{\frac{5\cdot 48^{d_{\mathcal{G},\gamma}}K_{\mathcal{G}}}{m}}C\cdot I(6). \quad (5)$$

The computation of the integral gives

$$I(t) = \frac{1}{2}\left[\sqrt{\ln(F(t))} + F(t)\frac{\sqrt{\pi}}{2}\operatorname{erfc}\left(\sqrt{\ln(F(t))}\right)\right]. \quad (6)$$

If T is a random variable following a standard normal distribution, then

$$\mathbb{P}(T \geqslant t) \leqslant \frac{1}{\sqrt{2\pi}}\frac{1}{t}e^{-\frac{t^2}{2}}.$$

A substitution of this classical tail bound in (6) provides us with:

$$I(t) \leqslant \frac{1}{2}\left[\sqrt{\ln(F(t))} + \sqrt{\frac{1}{4\ln(F(t))}}\right]. \quad (7)$$

A substitution of (7) (with $t = 6$) into (5) concludes the proof of (4).

Second case: $d_{\mathcal{G},\gamma} = 2$

$$R_m(\mathcal{F}_{\mathcal{G},\gamma}) \leqslant C\left(h(N) + 192\sqrt{\frac{5K_{\mathcal{G}}}{m}}\sum_{j\in\mathcal{J}}\frac{h(j)+h(j-1)}{h(j)}\sqrt{\ln\left(\frac{12M_{\mathcal{G}}}{h(j)}\right)}\right)$$

For $N = \left\lceil \frac{1}{2}\log_2(m)\right\rceil$, we set $h(j) = 2M_{\mathcal{G}}m^{-\frac{1}{2}}2^{-j+N}$. Then,

$$\begin{aligned} R_m(\mathcal{F}_{\mathcal{G},\gamma}) &\leqslant C\left(\frac{2M_{\mathcal{G}}}{\sqrt{m}} + 576\sqrt{\frac{5K_{\mathcal{G}}}{m}}\sum_{j=1}^{N}\sqrt{\ln\left(6\sqrt{m}\cdot 2^{j-N}\right)}\right) \\ &\leqslant C\left(\frac{2M_{\mathcal{G}}}{\sqrt{m}} + 576\sqrt{\frac{5K_{\mathcal{G}}}{m}}\sum_{j=1}^{N}\sqrt{\ln\left(6\sqrt{m}\right)}\right) \\ &\leqslant C\left(\frac{2M_{\mathcal{G}}}{\sqrt{m}} + 576\sqrt{\frac{5K_{\mathcal{G}}}{m}}\left\lceil\frac{1}{2}\log_2(m)\right\rceil\sqrt{\ln\left(6\sqrt{m}\right)}\right). \end{aligned}$$

Third case: $d_{\mathcal{G},\gamma} > 2$

For $N = \left\lceil \frac{d_{\mathcal{G},\gamma}-2}{2d_{\mathcal{G},\gamma}} \log_2 \left(\frac{m}{\log_2(m)} \right) \right\rceil$, let us set $h(j) = 2M_{\mathcal{G}} \left(\frac{\log_2(m)}{m} \right)^{\frac{1}{d_{\mathcal{G},\gamma}}} 2^{\frac{2}{d_{\mathcal{G},\gamma}-2}(-j+N)}$.
We then get

$$R_m(\mathcal{F}_{\mathcal{G},\gamma}) \leqslant 2M_{\mathcal{G}} C \left(\frac{\log_2(m)}{m} \right)^{\frac{1}{d_{\mathcal{G},\gamma}}} \left(1 + 4\left(1 + 2^{\frac{2}{d_{\mathcal{G},\gamma}-2}} \right) (2M_{\mathcal{G}})^{-\frac{d_{\mathcal{G},\gamma}}{2}} \sqrt{\frac{5 \cdot 48^{d_{\mathcal{G},\gamma}} K_{\mathcal{G}}}{\log_2(m)}} S\left(6 \left(\frac{m}{\log_2(m)} \right)^{\frac{1}{d_{\mathcal{G},\gamma}}} \right) \right),$$

where the series $S(t)$ is defined by:

$$\forall t \in \mathbb{R}_+^*,\ S(t) = \sum_{j=1}^{N} 2^{j-N} \sqrt{\ln\left(t \cdot 2^{\frac{2}{d_{\mathcal{G},\gamma}-2}(j-N)} \right)}.$$

$S(t)$ is trivially upper bounded as follows:

$$\begin{aligned} S(t) &\leqslant \sqrt{\ln(t)} \sum_{j=1}^{N} 2^{j-N} \\ &\leqslant 2\sqrt{\ln(t)}. \end{aligned} \tag{8}$$

4.2.2 Decomposition involving covering numbers

This time, we make another classical hypothesis: $m > C$. Applying in sequence Theorem 1, Lemma 2, Lemma 3, Theorem 1 in [22] and Hypothesis 1 gives:

$$\begin{aligned} R_m(\mathcal{F}_{\mathcal{G},\gamma}) &\leqslant h(N) + 2\sum_{j=1}^{N} (h(j) + h(j-1)) \sqrt{\frac{\ln\left(\mathcal{N}_2^{\text{int}}(h(j), \mathcal{F}_{\mathcal{G},\gamma}, m) \right)}{m}} \\ &\leqslant h(N) + 2\sqrt{\frac{C}{m}} \sum_{j=1}^{N} (h(j) + h(j-1)) \sqrt{\ln\left(\max_{1 \leqslant k \leqslant C} \mathcal{N}_2^{\text{int}}\left(\frac{h(j)}{\sqrt{C}}, \mathcal{G}_k, m \right) \right)} \\ &\leqslant h(N) + 4\sqrt{\frac{5 \cdot 48^{d_{\mathcal{G},\gamma}} K_{\mathcal{G}}}{m}} C^{\frac{d_{\mathcal{G},\gamma}+2}{4}} \sum_{j \in \mathcal{J}} \frac{h(j) + h(j-1)}{h(j)^{\frac{d_{\mathcal{G},\gamma}}{2}}} \sqrt{\ln\left(\frac{12M_{\mathcal{G}}\sqrt{C}}{h(j)} \right)} \end{aligned}$$

where $\mathcal{J} = \left\{ j \in [\![1; N]\!] :\ h(j) \leqslant 2M_{\mathcal{G}}\sqrt{C} \right\}$.

Theorem 4 *Let $\mathcal{G}$ be a class of functions satisfying Hypothesis 1. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 6. Suppose that $m > C$.*
If $d_{\mathcal{G}} \in (0,2)$, then

$$R_m(\mathcal{F}_{\mathcal{G},\gamma}) \leqslant 8 \frac{1 + 2^{\frac{2}{2-d_{\mathcal{G},\gamma}}}}{\sqrt{2(2-d_{\mathcal{G}})}} \gamma^{1-\frac{d_{\mathcal{G},\gamma}}{2}} \sqrt{\frac{5 \cdot 48^{d_{\mathcal{G},\gamma}} K_{\mathcal{G}}}{m}} C^{\frac{d_{\mathcal{G},\gamma}+2}{4}} \left[\sqrt{\ln\left(F\left(\frac{12M_{\mathcal{G}}\sqrt{C}}{\gamma} \right) \right)} + \sqrt{\frac{1}{4\ln\left(F\left(\frac{12M_{\mathcal{G}}\sqrt{C}}{\gamma} \right) \right)}} \right],$$

where F is the function defined in Theorem 3.
If $d_{\mathcal{G}} = 2$, then

$$R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \gamma\sqrt{\frac{C}{m}}\left(1 + 576\frac{1}{\gamma}\sqrt{5K_{\mathcal{G}}C}\left\lceil\frac{1}{2}\log_2\left(\frac{m}{C}\right)\right\rceil\sqrt{\ln\left(\frac{12M_{\mathcal{G}}\sqrt{m}}{\gamma}\right)}\right).$$

At last, if $d_{\mathcal{G},\gamma} > 2$, then

$$R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \gamma\sqrt{C}\left(\frac{C}{m}\right)^{\frac{1}{d_{\mathcal{G},\gamma}}}\left(1 + 8\left(1 + 2^{\frac{2}{d_{\mathcal{G},\gamma}-2}}\right)\left(\frac{1}{\gamma}\right)^{\frac{d_{\mathcal{G},\gamma}}{2}}\sqrt{5\cdot 48^{d_{\mathcal{G},\gamma}}K_{\mathcal{G}}}\sqrt{\ln\left(\frac{12M_{\mathcal{G}}}{\gamma}\left(\frac{m}{C}\right)^{\frac{1}{d_{\mathcal{G},\gamma}}}\right)}\right).$$

Proof
First case: $d_{\mathcal{G},\gamma} \in (0,2)$
This case is the only one for which the entropy integral exists. Setting for every $j \in \mathbb{N}$, $h\left(j\right) = \gamma 2^{-\frac{2}{2-d_{\mathcal{G},\gamma}}j}$, we obtain

$$\begin{aligned}
R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant\ & 8\left(1 + 2^{\frac{2}{2-d_{\mathcal{G},\gamma}}}\right)\gamma^{1-\frac{d_{\mathcal{G},\gamma}}{2}}\sqrt{\frac{5\cdot 48^{d_{\mathcal{G},\gamma}}K_{\mathcal{G}}}{m}}C^{\frac{d_{\mathcal{G},\gamma}+2}{4}}\int_0^{\frac{1}{2}}\sqrt{\ln\left(\frac{12M_{\mathcal{G}}\sqrt{C}}{\gamma\epsilon^{\frac{2}{2-d_{\mathcal{G},\gamma}}}}\right)}d\epsilon \\
=\ & 8\frac{\left(1 + 2^{\frac{2}{2-d_{\mathcal{G},\gamma}}}\right)\sqrt{2}}{\sqrt{2-d_{\mathcal{G},\gamma}}}\gamma^{1-\frac{d_{\mathcal{G},\gamma}}{2}}\sqrt{\frac{5\cdot 48^{d_{\mathcal{G},\gamma}}K_{\mathcal{G}}}{m}}C^{\frac{d_{\mathcal{G},\gamma}+2}{4}}\cdot I\left(\frac{12M_{\mathcal{G}}\sqrt{C}}{\gamma}\right).
\end{aligned}$$

The upper bound on the integral is provided by Inequality (7).

Second case: $d_{\mathcal{G},\gamma} = 2$

$$R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant h\left(N\right) + 192\sqrt{\frac{5K_{\mathcal{G}}}{m}}C\sum_{j\in\mathcal{J}}\frac{h\left(j\right)+h\left(j-1\right)}{h\left(j\right)}\sqrt{\ln\left(\frac{12M_{\mathcal{G}}\sqrt{C}}{h\left(j\right)}\right)}.$$

For $N = \left\lceil\frac{1}{2}\log_2\left(\frac{m}{C}\right)\right\rceil$, we set $h\left(j\right) = \gamma\sqrt{\frac{C}{m}}2^{-j+N}$.

$$\begin{aligned}
R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant\ & \gamma\sqrt{\frac{C}{m}} + 576\sqrt{\frac{5K_{\mathcal{G}}}{m}}C\sum_{j=1}^{N}\sqrt{\ln\left(\frac{12M_{\mathcal{G}}\sqrt{m}\cdot 2^{j-N}}{\gamma}\right)} \\
\leqslant\ & \gamma\sqrt{\frac{C}{m}} + 576\sqrt{\frac{5K_{\mathcal{G}}}{m}}C\cdot N\sqrt{\ln\left(\frac{12M_{\mathcal{G}}\sqrt{m}}{\gamma}\right)} \\
=\ & \gamma\sqrt{\frac{C}{m}} + 576\sqrt{\frac{5K_{\mathcal{G}}}{m}}C\left\lceil\frac{1}{2}\log_2\left(\frac{m}{C}\right)\right\rceil\sqrt{\ln\left(\frac{12M_{\mathcal{G}}\sqrt{m}}{\gamma}\right)}.
\end{aligned}$$

Third case: $d_{\mathcal{G},\gamma} > 2$
For $N = \left\lceil\frac{d_{\mathcal{G},\gamma}-2}{2d_{\mathcal{G},\gamma}}\log_2\left(\frac{m}{C}\right)\right\rceil$, let us set $h\left(j\right) = \gamma\sqrt{C}\left(\frac{C}{m}\right)^{\frac{1}{d_{\mathcal{G},\gamma}}}2^{\frac{2}{d_{\mathcal{G},\gamma}-2}(-j+N)}$.

We then get

$$R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \gamma\sqrt{C}\left(\frac{C}{m}\right)^{\frac{1}{d_{\mathcal{G},\gamma}}}\left(1+4\left(1+2^{\frac{2}{d_{\mathcal{G},\gamma}-2}}\right)\left(\frac{1}{\gamma}\right)^{\frac{d_{\mathcal{G},\gamma}}{2}}\sqrt{5\cdot 48^{d_{\mathcal{G},\gamma}}K_{\mathcal{G}}}S\left(\frac{12M_{\mathcal{G}}}{\gamma}\left(\frac{m}{C}\right)^{\frac{1}{d_{\mathcal{G},\gamma}}}\right)\right).$$

The upper bound on the series is provided by Inequality (8).

4.2.3 Decomposition involving scale-sensitive combinatorial dimensions

Applying in sequence Theorem 1, Lemma 3, Lemma 4 and Hypothesis 1 and setting $d_G\left(\epsilon\right) = \epsilon\text{-G-dim}\left(\mathcal{F}_{\mathcal{G}}\right)$ gives:

$$\begin{aligned}
R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) &\leqslant h\left(N\right) + 2\sum_{j=1}^{N}\left(h\left(j\right)+h\left(j-1\right)\right)\sqrt{\frac{\ln\left(\mathcal{N}_2^{\text{int}}\left(h\left(j\right),\mathcal{F}_{\mathcal{G},\gamma},m\right)\right)}{m}}\\
&\leqslant h\left(N\right) + 4\sqrt{\frac{5}{m}}\sum_{j\in\mathcal{J}}\left(h\left(j\right)+h\left(j-1\right)\right)\sqrt{d_G\left(\frac{h\left(j\right)}{48}\right)\ln\left(\frac{6\gamma}{h\left(j\right)}\right)}\\
&\leqslant h\left(N\right) + 4\sqrt{\frac{5\cdot 48^{d_{\mathcal{G},\gamma}}K_{\mathcal{G}}C^{d_{\mathcal{G},C}}}{m}}\sum_{j\in\mathcal{J}}\frac{h\left(j\right)+h\left(j-1\right)}{h\left(j\right)^{\frac{d_{\mathcal{G},\gamma}}{2}}}\sqrt{\ln\left(\frac{6\gamma}{h\left(j\right)}\right)}
\end{aligned}$$

where $\mathcal{J} = \left\{j \in [\![1;N]\!] :\ h\left(j\right) \leqslant \gamma\right\}$.

Theorem 5 *Let $\mathcal{G}$ be a class of functions satisfying Hypothesis 1. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 6.*
If $d_{\mathcal{G},\gamma} \in (0,2)$, then

$$R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant 8\frac{1+2^{\frac{2}{2-d_{\mathcal{G},\gamma}}}}{\sqrt{2\left(2-d_{\mathcal{G},\gamma}\right)}}\gamma^{1-\frac{d_{\mathcal{G},\gamma}}{2}}\sqrt{\frac{5\cdot 48^{d_{\mathcal{G},\gamma}}K_{\mathcal{G}}C^{d_{\mathcal{G},C}}}{m}}\left[\sqrt{\ln\left(F\left(6\right)\right)}+\sqrt{\frac{1}{4\ln\left(F\left(6\right)\right)}}\right],$$

where F is the function defined in Theorem 3.
If $d_{\mathcal{G},\gamma} = 2$, then

$$R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \frac{\gamma}{\sqrt{m}}\left(1+576\frac{1}{\gamma}\sqrt{5K_{\mathcal{G}}C^{d_{\mathcal{G},C}}}\left\lceil\frac{1}{2}\log_2\left(m\right)\right\rceil\sqrt{\ln\left(6\sqrt{m}\right)}\right).$$

At last, if $d_{\mathcal{G},\gamma} > 2$, then

$$R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \gamma\left(\frac{\log_2\left(m\right)}{m}\right)^{\frac{1}{d_{\mathcal{G},\gamma}}}\left(1+8\left(1+2^{\frac{2}{d_{\mathcal{G},\gamma}-2}}\right)\left(\frac{1}{\gamma}\right)^{\frac{d_{\mathcal{G},\gamma}}{2}}\sqrt{\frac{5\cdot 48^{d_{\mathcal{G},\gamma}}K_{\mathcal{G}}C^{d_{\mathcal{G},C}}}{\log_2\left(m\right)}}\sqrt{\ln\left(6\left(\frac{m}{\log_2\left(m\right)}\right)^{\frac{1}{d_{\mathcal{G},\gamma}}}\right)}\right).$$

Proof
First case: $d_{\mathcal{G},\gamma} \in (0,2)$
This case is the only one for which the entropy integral exists. Setting for every $j \in \mathbb{N}$, $h(j) = \gamma 2^{-\frac{2}{2-d_{\mathcal{G},\gamma}} j}$, we obtain

$$R_m(\mathcal{F}_{\mathcal{G},\gamma}) \leqslant 8\left(1+2^{\frac{2}{2-d_{\mathcal{G},\gamma}}}\right)\gamma^{1-\frac{d_{\mathcal{G},\gamma}}{2}}\sqrt{\frac{5\cdot 48^{d_{\mathcal{G},\gamma}}K_{\mathcal{G}}C^{d_{\mathcal{G},C}}}{m}}\int_0^{\frac{1}{2}}\sqrt{\ln\left(\frac{6}{\epsilon^{\frac{2}{2-d_{\mathcal{G},\gamma}}}}\right)}d\epsilon$$
$$= 8\frac{\left(1+2^{\frac{2}{2-d_{\mathcal{G},\gamma}}}\right)\sqrt{2}}{\sqrt{2-d_{\mathcal{G},\gamma}}}\gamma^{1-\frac{d_{\mathcal{G},\gamma}}{2}}\sqrt{\frac{5\cdot 48^{d_{\mathcal{G},\gamma}}K_{\mathcal{G}}C^{d_{\mathcal{G},C}}}{m}}\cdot I(6).$$

The upper bound on the integral is provided by Inequality (7).

Second case: $d_{\mathcal{G},\gamma} = 2$

$$R_m(\mathcal{F}_{\mathcal{G},\gamma}) \leqslant h(N) + 192\sqrt{\frac{5K_{\mathcal{G}}C^{d_{\mathcal{G},C}}}{m}}\sum_{j\in\mathcal{J}}\frac{h(j)+h(j-1)}{h(j)}\sqrt{\ln\left(\frac{6\gamma}{h(j)}\right)}.$$

For $N = \left\lceil \frac{1}{2}\log_2(m)\right\rceil$, we set $h(j) = \gamma m^{-\frac{1}{2}}2^{-j+N}$. Then,

$$R_m(\mathcal{F}_{\mathcal{G},\gamma}) \leqslant \frac{\gamma}{\sqrt{m}} + 576\sqrt{\frac{5K_{\mathcal{G}}C^{d_{\mathcal{G},C}}}{m}}\sum_{j=1}^{N}\sqrt{\ln\left(6\sqrt{m}\cdot 2^{j-N}\right)}$$
$$\leqslant \frac{\gamma}{\sqrt{m}} + 576\sqrt{\frac{5K_{\mathcal{G}}C^{d_{\mathcal{G},C}}}{m}}\left\lceil\frac{1}{2}\log_2(m)\right\rceil\sqrt{\ln\left(6\sqrt{m}\right)}.$$

Third case: $d_{\mathcal{G},\gamma} > 2$
For $N = \left\lceil \frac{d_{\mathcal{G},\gamma}-2}{2d_{\mathcal{G},\gamma}}\log_2\left(\frac{m}{\log_2(m)}\right)\right\rceil$, let us set $h(j) = \gamma\left(\frac{\log_2(m)}{m}\right)^{\frac{1}{d_{\mathcal{G},\gamma}}}2^{\frac{2}{d_{\mathcal{G},\gamma}-2}(-j+N)}$.
We then get

$$R_m(\mathcal{F}_{\mathcal{G},\gamma}) \leqslant \gamma\left(\frac{\log_2(m)}{m}\right)^{\frac{1}{d_{\mathcal{G},\gamma}}}\left(1+4\left(1+2^{\frac{2}{d_{\mathcal{G},\gamma}-2}}\right)\left(\frac{1}{\gamma}\right)^{\frac{d_{\mathcal{G},\gamma}}{2}}\sqrt{\frac{5\cdot 48^{d_{\mathcal{G},\gamma}}K_{\mathcal{G}}C^{d_{\mathcal{G},C}}}{\log_2(m)}}\,S\left(6\left(\frac{m}{\log_2(m)}\right)^{\frac{1}{d_{\mathcal{G},\gamma}}}\right)\right).$$

The upper bound on the series is provided by Inequality (8).

4.3 Discussion

A comparison of Theorems 3 to 5 is rich in lessons. First, the convergence rate varies with $d_{\mathcal{G},\gamma}$, not the level of the decomposition. Second, the dependence on C improves from linear to sublinear when the decomposition is postponed to the level of the covering numbers. Whether postponing further the decomposition brings something more utterly depends on the value of $d_{\mathcal{G},C}$. This implies that a favorable condition for this last option is a strong coupling among the outputs of the classifier. At last, the dependence on γ is better when performing the transition with scale-sensitive combinatorial dimensions: the gain is a factor $\sqrt{\ln(\gamma^{-1})}$.

5 New guaranteed risks based on the margin indicator loss function

All the bounds based on this loss function implement Pollard's combinatorial method (see Chapter 2 of [24]).

5.1 Basic supremum inequality

The basic supremum inequality is a multi-class extension of Lemma 4 in [3], with the first symmetrization being derived from the basic lemma of Section 4.5.1 in [27].

Theorem 6 (Theorem 2 in [12]) *Let $\mathcal{G}$ be a class of functions satisfying Definition 1. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 6. For a fixed $\gamma \in (0,1]$ and a fixed $\delta \in (0,1)$, with P^m-probability at least $1-\delta$, uniformly for every function $g \in \mathcal{G}$,*

$$L(g) \leqslant L_{\gamma,m}(g) + \sqrt{\frac{2}{m}\left(\ln\left(\mathcal{N}_\infty^{int}\left(\frac{\gamma}{2}, \mathcal{F}_{\mathcal{G},\gamma}, 2m\right)\right) + \ln\left(\frac{2}{\delta}\right)\right)} + \frac{1}{m}, \quad (9)$$

where the margin loss function is $\phi_{\infty,\gamma}$.

5.2 Upper bounds on the metric entropy

This time, there are only two levels at which the decomposition can take place.

5.2.1 Decomposition involving covering numbers

Applying in sequence Lemma 2, Lemma 3, Lemma 3.5 in [1] and Hypothesis 1 gives:

$$\ln\left(\mathcal{N}_\infty^{\text{int}}\left(\frac{\gamma}{2}, \mathcal{F}_{\mathcal{G},\gamma}, 2m\right)\right) \leqslant \frac{3}{2} K_{\mathcal{G}} C \left(\frac{8}{\gamma}\right)^{d_{\mathcal{G},\gamma}} \ln^2\left(\frac{128 M_{\mathcal{G}}^2 m}{\gamma^2}\right) + \ln\left(2^C\right). \quad (10)$$

The convergence rate of the guaranteed risk resulting from substituting the right-hand side of (10) into (9) is $\frac{\ln(m)}{\sqrt{m}}$. Its dependence on C is radical, which corresponds to the state of the art (see for instance [30,12]). At last, it varies with γ as a $O\left(\gamma^{-\frac{d_{\mathcal{G},\gamma}}{2}} \ln\left(\gamma^{-1}\right)\right)$.

5.2.2 Decomposition involving scale-sensitive combinatorial dimensions

Making use of Hypothesis 1, the decomposition performed with scale-sensitive combinatorial dimensions (Lemmas 3 and 8) gives:

$$\begin{aligned} \ln\left(\mathcal{N}_\infty^{\text{int}}\left(\frac{\gamma}{2}, \mathcal{F}_{\mathcal{G},\gamma}, 2m\right)\right) &\leqslant \frac{3}{2} \cdot \frac{\gamma}{8}\text{-G-dim}\left(\mathcal{F}_{\mathcal{G}}\right) \ln^2(16em) + \ln(2) \\ &\leqslant \frac{3}{2} K_{\mathcal{G}} C^{d_{\mathcal{G},C}} \left(\frac{8}{\gamma}\right)^{d_{\mathcal{G},\gamma}} \ln^2(16em) + \ln(2). \quad (11) \end{aligned}$$

The guaranteed risk associated with this alternative upper bound on the metric entropy exhibits the same convergence rate as the previous one. The radical dependence on C is replaced with a growth as a $O\left(C^{\frac{d_{\mathcal{G},C}}{2}}\right)$, which could prove to be worse for the main margin classifiers of the literature (we conjecture that they should satisfy $d_{\mathcal{G},C} > 1$). On the contrary, Formula (11) implies a growth with γ^{-1} as a $O\left(\gamma^{-\frac{d_{\mathcal{G},\gamma}}{2}}\right)$, i.e., a gain of a factor $\ln\left(\gamma^{-1}\right)$.

6 Application to M-SVMs

In the two preceding sections, guaranteed risks have been derived which involve the margin Graph dimension, and Hypothesis 1 has been used as a generic upper bound on this dimension. We now focus on a popular family of margin classifiers, the M-SVMs, to provide an illustration of the way this Hypothesis can be instantiated in practice. This calls for a definition of the underlying function class. We base the definition of the C-category SVMs on that of reproducing kernel Hilbert space (RKHS) [5] of $\mathbb{R}^C$-valued functions.

Definition 12 (RKHS of $\mathbb{R}^C$-valued functions $\mathbf{H}_{\kappa,C}$, after Section 6 of [29]) Let κ be a real-valued positive type function on $\mathcal{X}^2$ and let $\left(\mathbf{H}_\kappa, \langle\cdot,\cdot\rangle_{\mathbf{H}_\kappa}\right)$ be the corresponding RKHS. Let $\tilde{\kappa}$ be the real-valued positive type function on $\mathcal{Z}^2$ deduced from κ as follows:

$$\forall\left((x,k),(x',l)\right) \in \mathcal{Z}^2, \quad \tilde{\kappa}\left((x,k),(x',l)\right) = \delta_{k,l}\kappa\left(x,x'\right),$$

where δ is the Kronecker symbol. For every $(x,k) \in \mathcal{Z}$, let us define the $\mathbb{R}^C$-valued function $\tilde{\kappa}_{x,k}^{(C)}$ on $\mathcal{X}$ by the formula

$$\tilde{\kappa}_{x,k}^{(C)}(\cdot) = \left(\tilde{\kappa}\left((x,k),(\cdot,l)\right)\right)_{1\leqslant l\leqslant C}. \tag{12}$$

The RKHS of $\mathbb{R}^C$-valued functions at the basis of a C-category SVM whose kernel is κ, $\left(\mathbf{H}_{\kappa,C}, \langle\cdot,\cdot\rangle_{\mathbf{H}_{\kappa,C}}\right)$, consists of the linear manifold of all finite linear combinations of functions of the form (12) as (x,k) varies in $\mathcal{Z}$, and its closure with respect to the inner product

$$\forall\left((x,k),(x',l)\right) \in \mathcal{Z}^2, \quad \left\langle \tilde{\kappa}_{x,k}^{(C)}, \tilde{\kappa}_{x',l}^{(C)} \right\rangle_{\mathbf{H}_{\kappa,C}} = \tilde{\kappa}\left((x,k),(x',l)\right).$$

With the definition of this RKHS a hand, the specification of the function class at the basis of a C-category SVM rests on the introduction of a condition controlling the capacity through the characterization of a coupling among the outputs. We consider the standard condition, used for instance in [19].

Definition 13 (Function class $\mathcal{H}_\Lambda$) Let κ be a real-valued positive type function on $\mathcal{X}^2$ and let $\Lambda \in \mathbb{R}_+^*$. Let $\left(\mathbf{H}_{\kappa,C}, \langle\cdot,\cdot\rangle_{\mathbf{H}_{\kappa,C}}\right)$ be the RKHS of $\mathbb{R}^C$-valued functions spanned by κ according to Definition 12. Then the function

class $\mathcal{H}_\Lambda$ associated with the C-category SVM parameterized by (κ, Λ) is:

$$\mathcal{H}_\Lambda = \left\{ h = (h_k)_{1 \leqslant k \leqslant C} \in \mathbf{H}_{\kappa,C} : \sum_{k=1}^{C} h_k = 0_{\mathbf{H}_\kappa} \text{ and } \|h\|_{\mathbf{H}_{\kappa,C}} \leqslant \Lambda \right\}.$$

To upper bound the margin Graph dimension of the class $\mathcal{F}_{\mathcal{H}_\Lambda}$, we use an intermediate step involving another γ-Ψ-dimension: the margin Natarajan dimension γ-N-dim (Definition 18 in [13]). It is easy to notice that the proof of Lemma 15 in [13] still holds true if the strong Graph dimension is replaced with the strong dimension. A combination of Theorem 1 in [22] with this variant of Lemma 15 in [13] provides us with:

Lemma 9 *Let $\mathcal{G}$ be a class of functions satisfying Definition 1 and $\mathcal{F}_\mathcal{G}$ the class of functions deduced from $\mathcal{G}$ according to Definition 2. For $\epsilon \in (0, M_\mathcal{G}]$, let $d_N(\epsilon) = \epsilon$-N-dim$(\mathcal{F}_\mathcal{G})$. Then for $\epsilon \in (0, M_\mathcal{G}]$ and $n \in \mathbb{N}^*$,*

$$\mathcal{M}_2(\epsilon, \mathcal{F}_\mathcal{G}, n) \leqslant \left(\frac{12 M_\mathcal{G}}{\epsilon} \right)^{240 \log_2^{\alpha(C)}(F(C)) d_N^{\beta(C)}\left(\frac{\epsilon}{48}\right)},$$

where $F(C) = 4(C-1)$, $\alpha(C) = 2 + \frac{2}{2\ln(F(C))-1}$ and $\beta(C) = 1 + \frac{1}{4\ln(F(C))-2}$.

With Lemma 9 at hand, the following instantiation of Hypothesis 1 to the C-category SVMs is easy to establish.

Lemma 10 *For $\Lambda \in \mathbb{R}_+^*$, let $\mathcal{H}_\Lambda$ be a function class satisfying Definition 13. Suppose that for every $x \in \mathcal{X}$, κ_x belongs to the closed ball of radius $\Lambda_\mathcal{X}$ about the origin in $\mathbf{H}_\kappa$. Then, for every $\gamma \in (0, \Lambda\Lambda_\mathcal{X}]$,*

$$\gamma\text{-}G\text{-}dim(\mathcal{F}_{\mathcal{H}_\Lambda}) \leqslant 7114 C \log_2^{\alpha(C)}(F(C)) \left(\frac{24\Lambda\Lambda_\mathcal{X}}{\gamma} \right)^{2\beta(C)} \ln\left(\frac{12\Lambda\Lambda_\mathcal{X}}{\gamma} \right),$$

with the functions F, α and β being defined in Lemma 9.

Proof The proof results from applying in sequence Proposition 5 in [13] (with $p = 2$), Lemma 9 (above) and Lemma 10 in [13]. This gives

$$\begin{aligned}
\forall \gamma \in (0, \Lambda\Lambda_\mathcal{X}],\ \gamma\text{-G-dim}(\mathcal{F}_{\mathcal{H}_\Lambda}) &\leqslant \frac{16}{\ln(2)} \ln\left(\mathcal{M}_2(\gamma, \mathcal{F}_{\mathcal{H}_\Lambda}, \gamma\text{-dim}(\mathcal{F}_{\mathcal{H}_\Lambda}))\right) \\
&\leqslant \frac{3840}{\ln(2)} \log_2^{\alpha(C)}(F(C)) d_N^{\beta(C)}\left(\frac{\gamma}{48}\right) \ln\left(\frac{12\Lambda\Lambda_\mathcal{X}}{\gamma}\right) \\
&\leqslant \frac{3840}{\ln(2)} C^{\beta(C)} \log_2^{\alpha(C)}(F(C)) \left(\frac{24\Lambda\Lambda_\mathcal{X}}{\gamma}\right)^{2\beta(C)} \ln\left(\frac{12\Lambda\Lambda_\mathcal{X}}{\gamma}\right) \\
&\leqslant 7114 C \log_2^{\alpha(C)}(F(C)) \left(\frac{24\Lambda\Lambda_\mathcal{X}}{\gamma}\right)^{2\beta(C)} \ln\left(\frac{12\Lambda\Lambda_\mathcal{X}}{\gamma}\right).
\end{aligned}$$

This upper bound on γ-G-dim$(\mathcal{F}_{\mathcal{H}_\Lambda})$ compares with that obtained with the straightforward approach (without involving the margin Natarajan dimension), which consists in combining Lemma 8 in [13] with Theorem 4.6 in [4]. It exhibits a better dependence on C, a $O(C \ln(C))$, which is mainly due to the fact that Lemma 10 in [13] takes efficiently into account the coupling among the outputs.

7 Conclusions and ongoing research

This article has addressed the question of the level at which to perform the transition from the multi-class case to the bi-class one when deriving a guaranteed risk for margin multi-category classifiers. The comparative study was performed for the parameterized truncated hinge loss function and the margin indicator loss function. In both cases, all the possible decompositions (transitions from either $\mathcal{F}_{\mathcal{G},\gamma}$ or $\mathcal{F}_{\mathcal{G}}$ to a function class including the classes $\mathcal{G}_k$) have been considered. The following conclusions can be drawn from the corresponding bounds. First, the convergence rate depends on the choice of the margin loss function and the behaviour of the scale-sensitive combinatorial dimensions involved, but not on the level of the decomposition. Second, whatever the margin loss function, one can always exhibit a bound with sublinear dependence on C. The margin Graph dimension appears as an efficient tool to achieve this result, in the case when there exists a strong coupling among the outputs of the classifier. Third, the dependence on γ also benefits from a late decomposition.

Our work continues in two directions: taking into account the nature of the *learner/learning algorithm* [7] and investigating the instantiations of Hypothesis 1 obtained by deriving upper bounds on γ-Ψ-dimensions of other margin classifiers.

Acknowledgements This work was partly funded by a CNRS research grant (PEPS). Conflict of Interest: The author declares that he has no conflict of interest.

References

1. Alon, N., Ben-David, S., Cesa-Bianchi, N., Haussler, D.: Scale-sensitive dimensions, uniform convergence, and learnability. Journal of the ACM **44**(4), 615–631 (1997)
2. Anthony, M., Bartlett, P.: Neural Network Learning: Theoretical Foundations. Cambridge University Press, Cambridge (1999)
3. Bartlett, P.: The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. IEEE Transactions on Information Theory **44**(2), 525–536 (1998)
4. Bartlett, P., Shawe-Taylor, J.: Generalization performance of support vector machines and other pattern classifiers. In: B. Schölkopf, C. Burges, A. Smola (eds.) Advances in Kernel Methods - Support Vector Learning, chap. 4, pp. 43–54. The MIT Press, Cambridge, MA (1999)
5. Berlinet, A., Thomas-Agnan, C.: Reproducing Kernel Hilbert Spaces in Probability and Statistics. Kluwer Academic Publishers, Boston (2004)
6. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995)
7. Daniely, A., Sabato, S., Ben-David, S., Shalev-Shwartz, S.: Multiclass learnability and the ERM principle. In: COLT'11, pp. 207–232 (2011)
8. Doğan, U., Glasmachers, T., Igel, C.: A unified view on multi-class support vector classification. Journal of Machine Learning Research **17(45)**, 1–32 (2016)
9. Dudley, R., Giné, E., Zinn, J.: Uniform and universal Glivenko-Cantelli classes. Journal of Theoretical Probability **4**(3), 485–510 (1991)
10. Guermeur, Y.: VC theory of large margin multi-category classifiers. Journal of Machine Learning Research **8**, 2551–2594 (2007)

11. Guermeur, Y.: A generic model of multi-class support vector machine. International Journal of Intelligent Information and Database Systems **6**(6), 555–577 (2012)
12. Guermeur, Y.: L_p-norm Sauer-Shelah lemma for margin multi-category classifiers. Journal of Computer and System Sciences **89**, 450–473 (2017)
13. Guermeur, Y.: Combinatorial and structural results for γ-ψ-dimensions. Tech. rep., arXiv:1809.07310 (2018)
14. Kearns, M., Schapire, R.: Efficient distribution-free learning of probabilistic concepts. Journal of Computer and System Sciences **48**(3), 464–497 (1994)
15. Kearns, M., Schapire, R., Sellie, L.: Toward efficient agnostic learning. Machine Learning **17**(2-3), 115–141 (1994)
16. Kolmogorov, A., Tihomirov, V.: ϵ-entropy and ϵ-capacity of sets in functional spaces. American Mathematical Society Translations, series 2 **17**, 277–364 (1961)
17. Kontorovich, A., Weiss, R.: Maximum margin muliclass nearest neighbors. In: ICML'14 (2014)
18. Kuznetsov, V., Mohri, M., Syed, U.: Multi-class deep boosting. In: NIPS 27, pp. 2501–2509 (2014)
19. Lei, Y., Doğan, U., Binder, A., Kloft, M.: Multi-class SVMs: From tighter data-dependent generalization bounds to novel algorithms. In: NIPS 28, pp. 2026–2034 (2015)
20. Maurer, A.: A vector-contraction inequality for Rademacher complexities. In: ALT'16, pp. 3–17 (2016)
21. Mendelson, S.: A few notes on statistical learning theory. In: S. Mendelson, A. Smola (eds.) Advanced Lectures on Machine Learning, chap. 1, pp. 1–40. Springer-Verlag, Berlin, Heidelberg, New York (2003)
22. Mendelson, S., Vershynin, R.: Entropy and the combinatorial dimension. Inventiones mathematicae **152**, 37–55 (2003)
23. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. The MIT Press, Cambridge, MA (2012)
24. Pollard, D.: Convergence of Stochastic Processes. Springer-Verlag, New York (1984)
25. Talagrand, M.: Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems. Springer-Verlag, Berlin Heidelberg (2014)
26. van der Vaart, A., Wellner, J.: Weak Convergence and Empirical Processes, With Applications to Statistics. Springer Series in Statistics. Springer-Verlag, New York (1996)
27. Vapnik, V.: Statistical Learning Theory. John Wiley & Sons, Inc., New York (1998)
28. Vapnik, V., Chervonenkis, A.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications **XVI**(2), 264–280 (1971)
29. Wahba, G.: Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In: M. Casdagli, S. Eubank (eds.) Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity, vol. XII, pp. 95–112. Addison-Wesley (1992)
30. Zhang, T.: Statistical analysis of some multi-category large margin classification methods. Journal of Machine Learning Research **5**, 1225–1251 (2004)